Spectral Properties of Random Matrices for Stochastic Block Model

Konstantin Avrachenkov, Laura Cottatellucci, Arun Kadavankandy

▶ **To cite this version:**

Konstantin Avrachenkov, Laura Cottatellucci, Arun Kadavankandy. Spectral Properties of Random Matrices for Stochastic Block Model. [Research Report] RR-8703, INRIA Sophia-Antipolis, France; INRIA. 2015. hal-01142944

HAL Id: hal-01142944

https://inria.hal.science/hal-01142944

Submitted on 16 Apr 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

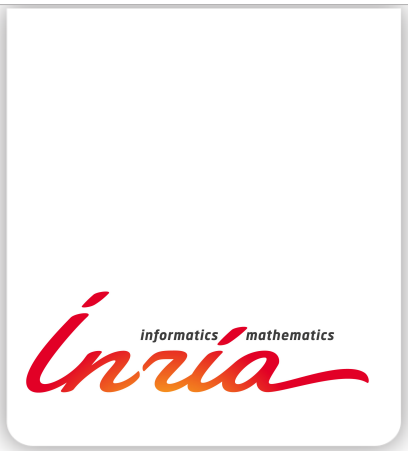
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Spectral Properties of Random Matrices for Stochastic Block Model

Konstantin Avrachenkov, Laura Cottatellucci ,
Arun Kadavankandy

RESEARCH REPORT
N° 8703
Apr 2015
Project-Team Maestro

ISSN 0249-6399 ISRN INRIA/RR--8703--FR+ENG



Spectral Properties of Random Matrices for Stochastic Block Model

Konstantin Avrachenkov, Laura Cottatellucci *,
Arun Kadavankandy

Project-Team Maestro

Research Report n° 8703 — Apr 2015 — 23 pages

Abstract: We consider an extension of Erdős-Rényi graph known in literature as Stochastic Block Model (SBM). We analyze the limiting empirical distribution of the eigenvalues of the adjacency matrix of SBM. We derive a fixed point equation for the Stieltjes transform of the limiting eigenvalue empirical distribution function (e.d.f.), concentration results on both the support of the limiting e.s.f. and the extremal eigenvalues outside the support of the limiting e.d.f. Additionally, we derive analogous results for the normalized Laplacian matrix and discuss potential applications of the general results in epidemics and random walks.

Key-words: Spectral Graph Theory, Random Graphs, Random Matrices,Stochastic Block Model

* Eurecom, France

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Carectéristiques Spectrales des Matrices Aléatoires pour Stochastic Block Model

Résumé : Nous considérons une modele des graphes aleotoires connue dans la littérature par son nom de «Stochastic Block Model(SBM) ». Nous analysons la distribution empirique des valeurs propres de la matrice d'adjacence du SBM. Nous obtenons une equation de point fixe pour la transformation de Stieltjes de cette fonction, et nous montrons l'existence d'un trou spectral dans la distribution. En outre, nous derivons les résultats similaires pour la matrice de Laplace normalisé et nous discoutons des applications potentielles des résultats généraux dans les études d'épidémies et marches aléatoires.

Mots-clés : Theorie spectrale de graphes, Matrices Aléatoires, Graphes Aléatoires, Stochastic Block Model

1 Introduction

Systems consisting of a huge number of interacting entities are often represented by complex networks to capture the essence of their interactions. They are used to model systems from the most disparate fields: interactions among atoms in matter, biological molecules, nodes of the Internet networks, documents in the web, social networks, connectivity of users in a wireless network, etc. Due to the typical gigantic dimensions of the systems targeted in this field, it is essential to gain understanding and master the system via few fundamental parameters which are able to characterize the macroscopic features of the system.

One important approach to study complex networks is based on the theory of random graphs. The first natural random graph model of complex networks is Erdős-Rényi graph [12] where edges between nodes appear with equal probabilities. This model has many appealing analytical properties but unfortunately does not model important properties of many real complex networks. In particular, the Erdős-Rényi graph fails in describing community structures in complex networks. To mitigate this shortcoming the more refined Stochastic Block Model (SBM) has been introduced, first studied in [9] to the best of our knowledge. In SBM, the nodes are divided into subsets (communities) such that nodes within a given community have a link between them with a higher probability than nodes from different communities.

Random graphs can generate a variety of random matrices, e.g. adjacency matrix, standard Laplacian, normalized Laplacian. The spectral properties of those random matrices are fundamental tools to predict and analyze the complex network behavior, the performance of random algorithms, e.g. searching algorithms, on the network, but also to design algorithms. For example, the convergence properties of a random walk on a graph are dependent on the gap between the first and the second largest eigenvalues of its normalized Laplacian [8]. This is of particular significance in random search algorithms deployed widely for information retrieval in big data systems. The community detection problem relies on the properties of several of the largest eigenvalues of the adjacency matrix and their corresponding eigenvectors, e.g. [20]. The cost of epidemic spread is characterized by the spectral properties of the adjacency matrix [7].

Recently, there has been a stream of works on SBM in community detection problems [10,17, 18]. In [10], the authors investigate detectability of communities in a SBM graph by analyzing phase transition in the spectrum of the adjacency matrix using methods from statistical physics. The authors of [17] analyze a similar problem in the context of labeled stochastic matrices with two communities and provide theoretical evidence for detectability thresholds in [10]. There, the nodes are randomly categorized into communities, with symmetric probabilities and the goal is to detect the correct community to which a node belongs. All these works focus on the case of diluted graphs, i.e. when the expected degrees are bounded irrespectively of the network size. In [18] the authors also study the detectability problem in the two community case. There, the scaling law of probabilities is not clearly defined. The case of dense graphs has been studied in [13] for M communities.

In this contribution, we analyze the limiting empirical distribution of the eigenvalues of the adjacency matrix of SBM. We find that the Stieltjes transform of the limiting distribution satisfies a fixed point equation and provide an explicit expression in the case of symmetric communities. Furthermore, we obtain tight bound on the support of the asymptotic spectrum, and concentration bounds on the extremal eigenvalues. Additionally, we derive the asymptotic spectrum of the normalized Laplacian matrix in the dense regime of edge probabilities and subsequently discuss potential applications of the general results in epidemics and random walks.

In comparison to the previous works on SBM [10, 13, 17, 18] we consider the dense and sparse regimes when the expected degrees scale not slower than $\log^c(n)$, for some $c \geq 4$.

In contrast to the classical Erdős-Rényi graph, the limiting eigenvalue e.d.f. of SBM adjacency

matrix does not follow the semicircular law anymore and there are M largest eigenvalues outside the support of the e.d.f. The concentration results on the extreme eigenvalues become more complex. Specifically, the result in [21], if directly applied, leads to a weaker bound on the edge of the e.d.f. support than the one we derive here. The authors of [10, 17, 18] consider only the case of two communities and their work does not include the study the limiting e.d.f. In [13] a general M community SBM is analyzed in the quite dense regime, i.e. when the average degree is of the order $n/\log(n)$.

2 Mathematical Notation and Definitions

Throughout this paper, 1_C is a C-dimensional column vector with unit entries and $J_C = 1_C 1_C^T$ is a square matrix whose entries are all equal to 1. The notations $f_n = O(g_n)$ and $f_n = o(g_n)$ mean that $\lim_{n\to\infty} \frac{|f_n|}{|g_n|} \leq c$ for some constant $c > 0$ and $\lim_{n\to\infty} \frac{f_n}{g_n} = 0$, respectively. We say that $f(x)$ dominates $g(x)$ asymptotically and we write $f(n) \in \omega(g(n))$ if asymptotically for $n \to +\infty$, $|f(n)| > k|g(n)|$ for any constant $k > 0$. By $\delta_y(x)$ we denote the Kronecker delta function equal to one when $x = y$ and zero everywhere else and the ceiling function that maps a real number to the smallest following integer is denoted by $\lceil\cdot\rceil$. The indicator function of a subset $\mathcal{A}$ of a set $\mathcal{X}$ is denoted by $\chi_{\mathcal{A}}(x)$ and, for any $x \in \mathcal{X}$, $\chi_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and $\chi_{\mathcal{A}}(x) = 0$ otherwise. The field of complex numbers is represented by $\mathbb{C}$, the operator $\Im(\cdot)$ maps a complex number onto its imaginary part and $\mathbb{C}^+$ denotes the open half space of complex numbers with nonnegative imaginary part, i.e., $z \in \mathbb{C}^+$, iif $\Im(z) > 0$. Given a random variable X and a distribution $\mathcal{D}$, the notation $X \sim \mathcal{D}$ indicates that the random variable X follows the probability distribution $\mathcal{D}$ and $P_X(x)$ denotes its cumulative distribution function.

Given an $n \times n$ Hermitian matrix H, we index its eigenvalues in nonincreasing order and denote by $\lambda_i(H)$ the i-th eigenvalue of H, i.e. $\lambda_1(H) \geq \lambda_2(H) \geq \lambda_3(H) ... \geq \lambda_n(H)$. The operator $\|\cdot\|_2$ denotes the Euclidian norm of a vector when the argument is a vector and spectral norm[1] when the argument is a matrix, i.e.

$$\|H\|_2 = \sup_{\|x\|_2 \leq 1, \|y\|_2 \leq 1} x^T H y. \tag{1}$$

The empirical spectral distribution (e.s.d.) of a Hermitian matrix H, is defined as:

$$F^H(x, n) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(H)}(x) \tag{2}$$

An important tool that is used in Random Matrix Analysis is the Stieltjes transform. It is extensively used to study various properties of the limiting spectral distribution such as the limiting shape and speed of convergence. Refer to [2] for more details. The Stieltjes transform $s_F(z)$ of a probability distribution $F(x)$ is defined as the Riemann-Stieltjes integral

$$s_F(z) = \int \frac{dF(x)}{x - z} \tag{3}$$

for $z \in \mathbb{C}^+$.

[1] The spectral norm coincides also with the induced 2-norm.

3 Stochastic Block Model and its Representations

We consider a complex network with n nodes and M communities Ω_m, for $m = 1, \ldots, M$, of equal sizes $K = n/M$, which is assumed to be an integer. This complex network is described by an undirected random graph referred to as Stochastic Block Model (SBM) with M blocks, one for each community. If two nodes belong to different communities, then there is an edge between them with probability $p_0(n)$. Given two nodes belonging to the same community Ω_m, there exists an edge between them with probability $p_m(n)$, $1 \leq m \leq M$. Throughout this paper, for the sake of conciseness, we adopt the short notation p_m for the probabilities $p_m(n)$, keeping in mind that the dependence on n is implicit. For a random graph as defined above, we can define a number of related random matrices whose spectral characteristics are relevant to capture related properties of the network. In this work we focus on two classes of random matrices for the SBM: the adjacency matrix and the normalized Laplacian matrix.

SBM adjacency matrix A

Without loss of generality, we assume that nodes belonging to the same community are clustered together and ordered from community 1 to community M, i.e. node i belongs to community Ω_m if $\left\lceil \frac{i}{K} \right\rceil = m$. The SBM adjacency matrix A is a symmetric matrix and its element A_{ij} is a Bernoulli variable with parameter p_m, $m = 1, \ldots M$, if the corresponding nodes i and j belong to the community Ω_m, i.e. $\left\lceil \frac{i}{K} \right\rceil = \left\lceil \frac{j}{K} \right\rceil = m$, and with parameter p_0 otherwise. Let us denote by $\mathcal{B}(p_m)$ a Bernoulli probability distribution with parameter p_m, then

$$\begin{cases} A_{ij} = A_{ji} \ \sim \ \mathcal{B}(p_m), & \text{if } i, j \in \Omega_m \\ A_{ij} = A_{ji} \ \sim \ \mathcal{B}(p_0), & \text{if } i \in \Omega_\ell \text{ and } j \in \Omega_m, \ell \neq m. \end{cases} \tag{4}$$

We implicitly assume that the diagonal elements of the matrix A are randomly distributed according to a given Bernoulli probability distribution. This corresponds to the assumption that the random graph has cycles of unit length with a certain probability. There are definitions of complex networks that do not admit cycles of unit length, which corresponds to matrices A with diagonal elements deterministically equal to zero. It is worth noting that the results on the asymptotic spectrums of adjacency matrices in this contribution hold for both these definitions under the assumptions made.

For further studies, it is convenient to normalize the matrix A by a scaling factor[2] $\gamma(n)$ in general depending on n such that the support of the limiting eigenvalue distribution function stay finite and positive. Then, we consider the normalized SBM adjacency matrix $\widehat{A} = \gamma(n)A$ and we express it as the sum of a deterministic matrix $\overline{A}$ equal to its expectation and a random matrix with zero mean random entries $\widetilde{A}$, i.e.

$$\widehat{A} = \overline{A} + \widetilde{A}. \tag{5}$$

Consistently with the definitions in (4) and (5), $\overline{A}$, the expectation of matrix A, is a finite rank matrix of the following form:

$$\overline{A} = P \otimes J_K \tag{6}$$

[2]We use the short notation γ when it is not necessary to emphasize the dependency on n.

being P the $M \times M$

$$P = \gamma(n) \begin{pmatrix} p_1 & p_0 & \dots & p_0 \\ p_0 & p_2 & \ddots & p_0 \\ \dots & & \ddots & \dots \\ p_0 & \dots & \dots & p_M \end{pmatrix}. \tag{7}$$

In general, for $p_m \neq p_0$ and $m = 1, \dots, M$, the matrix P has rank M and thus, also $\overline{A}$ has rank M.

The random centered SBM adjacency matrix is also a symmetric matrix whose elements follow the distributions

$$\mathcal{C}(p_m, \gamma) = \begin{cases} \gamma(1-p_m), & \text{w.p.} \quad p_m; \\ -\gamma p_m, & \text{w.p.} \quad 1-p_m; \end{cases} \qquad m = 0, 1, \dots, M, \tag{8}$$

having zero mean and variance $\sigma_m^2 = \gamma^2(1-p_m)p_m$. Consistently, with the definitions in (4) and (5)

$$\begin{cases} \widetilde{A}_{ij} = \widetilde{A}_{ji} \sim \mathcal{C}(p_m, \gamma) & \text{if } i, j \in \Omega_m \\ \widetilde{A}_{ij} = \widetilde{A}_{ji} \sim \mathcal{C}(p_0, \gamma) & \text{if } i \in \Omega_\ell \text{ and } m \in \Omega_m \text{ with } \ell \neq m. \end{cases} \tag{9}$$

It is worth noting that the entries of this random matrix depend of the matrix size.

Random SBM normalized Laplacian matrix $\mathcal{L}$

Let us define the random variable

$$D_i = \sum_{j=1}^{n} A_{ij} \tag{10}$$

corresponding to the degree of node i. Then, the symmetric SBM normalized Laplacian matrix $\mathcal{L}$ is defined as

$$\mathcal{L}_{ij} = \mathcal{L}_{ji} = \begin{cases} 1 - \dfrac{A_{ii}}{D_i}, & \text{if } i = j; \\ -\dfrac{A_{ij}}{\sqrt{D_i D_j}}, & \text{otherwise.} \end{cases} \tag{11}$$

4 Useful Existing Results

4.1 Erdős Rényi Graphs and Wigner matrices

A random graph where all the pairs of nodes have equal probability $p(n)$ of having an edge, independently of the presence of other edges is well-known as Erdős-Rényi (ER) graph. It is straightforward to verify that an SBM graph with $M = 1$, corresponding to a complex network with a single community, reduces to an ER graph. As for random SBM graphs, we can consider representations of random ER graphs by classes of random matrices. In this paper, we focus on random ER adjacency matrices A^{ER}. The upper diagonal elements of the Hermitian matrix A^{ER} are independent and identically distributed (iid) according to $\mathcal{B}(p(n))$, a Bernoulli distribution with parameter $p(n)$, i.e. $A_{ij}^{\text{ER}} = A_{ji}^{\text{ER}} \sim \mathcal{B}(p(n))$. As for random SBM adjacency matrices,

we consider a matrix $\widehat{A}^{\mathrm{ER}}$ normalized by the scalar $\gamma(n) = (\sqrt{np(n)(1-p(n))})^{-1}$, i.e. $\widehat{A}^{\mathrm{ER}} = \gamma(n)A^{\mathrm{ER}}$, and decompose it as

$$\widehat{A}^{\mathrm{ER}} = \overline{A}^{\mathrm{ER}} + \widetilde{A}^{\mathrm{ER}},$$

where $\overline{A}^{\mathrm{ER}} = \gamma(n)p(n)J_n$ and the centered ER adjacency matrix is given by

$$\widetilde{A}^{\mathrm{ER}}_{ij} = \widetilde{A}^{\mathrm{ER}}_{ji} \sim \mathcal{C}(p(n), (\sqrt{np(n)(1-p(n))})^{-1}) \ \forall i \geq j, i = 1, \dots n. \tag{12}$$

The parameter $p(n)$ depends on the network size n and, thus, also the average degree of a node i

$$d_{i,\mathrm{av}} = \mathrm{E}\left\{\sum_j A_{i,j}\right\} = np(n). \tag{13}$$

Based on the average node degree $d_{i,\mathrm{av}}$, the ER graphs are classified as *dense*, if $d_{i,\mathrm{av}} = O(n)$, *sparse* if $d_{i,\mathrm{av}} = o(n)$ and $d_{i,\mathrm{av}} \to \infty$, and *diluted* if $d_{i,\mathrm{av}} = O(1)$ [6] .

Closely related to the centered ER adjacency matrix is the Wigner matrix defined as a Hermitian matrix W whose upper diagonal entries are zero mean independent random variables with variance equal to σ^2.

It is worth noting that the centered ER adjacency matrix is a special case of Wigner matrices with a well defined distribution $\mathcal{C}(p(n), (\sqrt{np(n)(1-p(n))})^{-1}))$ equal for all the entries.

The properties of the eigenvalue spectrum of both Wigner matrices and ER adjacency matrices have been intensively studied. In this section, we recall the results on the limiting spectral distributions and the spectral norms of these random matrices. Defined the empirical spectral distribution (e.s.d.) of a Hermitian matrix H of size n as in (2), the limiting spectral distribution $F^H(x)$ is the deterministic limiting distribution, if it exists, of the random e.d.f. as the size of the matrix H tends to infinity. The spectral norm of a matrix is defined in (1).

4.2 Limiting e.s.d. of Centered ER Adjacency Matrices

A key role in the convergence of the e.s.d. of large random Hermitian matrices is played by the following assumption.

Assumption 1 *[15, 16, Chapter 1] The Hermitian matrix H with zero mean independent upper diagonal entries H_{ij} of variance σ_{ij}^2 such that $\lim_{n\to+\infty} \sup_{i,j=1,\dots n} \sigma_{ij}^2 = 0$ satisfies the Lindeberg's condition, i.e. for any $\delta > 0$*

$$\lim_{n\to\infty} \max_{i=1,\dots n} \sum_{j=1}^{n} \int_{|x|>\delta} x^2 \mathrm{d}P_{H_{ij}}(x) = 0. \tag{14}$$

This assumption essentially implies that the tails of the distributions characterizing the random variables H_{ij} diminish as $n \to \infty$. Under such an assumption, the sequence of the e.s.d. converges weakly to a limiting eigenvalue distribution in the almost sure sense as stated by the following proposition.

Proposition 1 *[15, 16, Chapter 1] The Wigner matrix W with zero mean independent random entries W_{ij} satisfies Assumption (1) and additionally, all the equal variances satisfy $\sigma_{i,j}^2 = \sigma^2/n$ with $0 < \sigma^2 < +\infty$. Then, the sequence of the e.s.d. converges weakly to a the Wigner semicircle law in the almost sure sense, i.e. for any bounded continuous function f*

$$\int f(x)F^{W_n}(x)dx \xrightarrow{a.s.} \int f(x)\mu_{sc}(x,\sigma^2)dx,$$

where $F^{W_n}(x)$ denotes an e.d.f. of the Wigner matrix of size n and $\mu_{sc}(x, \sigma^2)$ is the Wigner semicircular distribution with parameter σ^2 given by

$$\mu_{sc}(x, \sigma^2) = \frac{1}{2\pi\sigma^2}\sqrt{(4\sigma^2 - x^2)_+}. \tag{15}$$

This result can be immediately specialized to normalized centered ER adjacency matrices $\widehat{A}^{\mathrm{ER}}$. Since for the matrix $\widetilde{A}^{\mathrm{ER}}$ it holds $\sigma_{ij}^2 = n^{-1}$, for $i, j = 1, \dots n$, the conditions of Proposition (1) are satisfied if the limit (14) holds, i.e. if for any $\tau > 0$

$$\lim_{n\to+\infty}(1-p)\chi\left(1 - p \geq \tau\sqrt{np(1-p)}\right) + p\chi\left(p \geq \tau\sqrt{np(1-p)}\right) = 0. \tag{16}$$

It is straightforward to verify that this condition is equivalent to the condition $p \geq (\tau^2 n + 1)^{-1}$ for any $\tau > 0$. Then, we can state the following corollary.

Corollary 1 *Let us consider a normalized centered ER adjacency matrices $\widetilde{A}^{ER}$ with $p(n) \in \omega(n^{-1})$ as $n \to \infty$. Then, the sequence of the e.s.d. converges weakly to a the Wigner semicircle law in the almost sure sense, i.e. for any bounded continuous function f*

$$\int f(x) F^{\widetilde{A}^{ER}}(x)dx \xrightarrow{a.s.} \int f(x)\mu_{sc}(x, 1)dx.$$

Then, whether the e.s.d. of a centered ER adjacency matrix converges to a semi-circle distribution depends the convergence rate of $p(n)$ to zero as $n \to +\infty$. It is difficult to state any results on the limiting eigenvalue e.d.f. when $p(n) = \frac{c}{n}$ because for this probability, Assumption (1) does not hold. It is known that for diluted graphs, there is no explicit expression for the limiting eigenvalue e.d.f. but it displays atoms [6]. For this reason in the following we limit our attention to probabilities $p(n)$ such that $p(n) \in \omega(n^{-1})$, i.e. for the dense and sparse regime. This is tantamount to stating that $\sqrt{np_n(1-p_n)} \to \infty$ [11].

4.3 Spectral Norm of the Centered ER Adjacency Matrix

Let us observe that, if the multiplicity of an eigenvalue does not scale with n, the definition of the e.d.f. implies that, in the limit for $n \to +\infty$, the eigenvalue e.s.d. is not able to capture the existence of this eigenvalue in the spectrum matrix. Then, Corollary (1) can only provide a lower bound of the spectral norm of the normalized centered ER adjacency matrix $\widetilde{A}^{\mathrm{ER}}$. Hence, it is important to find an upper bound on the spectral norm of $\widetilde{A}^{\mathrm{ER}}$ to better understand its spectral properties. The following result comes in handy.

Lemma 1 *[21] Let W be a Wigner matrix with independent random elements W_{ij}, $i, j = 1, \dots n$ having zero mean and variance at most $\sigma^2(n)$. If the entries are bounded by $K(n)$ and there exist a constant C' such that $\sigma(n) \geq C' n^{-1/2} K(n) \log^2(n)$, then there exists a constant C such that almost surely*

$$\|W\|_2 \leq 2\sigma(n)\sqrt{n} + C(K(n)\sigma(n))^{1/2} n^{1/4} \log(n). \tag{17}$$

By applying Lemma (1) to the normalized centered adjacency matrix $\widetilde{A}^{\mathrm{ER}}$ we obtain the following concentration result.

COROLLARY 2 *Let us consider the normalized centered adjacency matrix $\widetilde{A}^{ER}$. If the probability $p(n)$ satisfies the inequality $p(n) \geq C' \log^4(n) n^{-1}$ for some constant $C' > 0$, then there exists a constant $C > 0$ such that almost surely*

$$\|\widetilde{A}^{ER}\|_2 \leq 2 + C \sqrt[4]{\frac{1-p(n)}{np(n)}} \log n. \tag{18}$$

Proof: From the definition of $\widetilde{A}^{\mathrm{ER}}$ it results $\sigma = n^{-1/2}$. Then, condition $\sigma \geq C^* n^{-1/2} K \log^2(n)$ in Lemma 1 implies $K \leq (C^* \log^2 n)^{-1}$. Additionally, the bound on the elements $\widetilde{A}^{\mathrm{ER}}_{ij}$ implies $\dfrac{1-p}{\sqrt{n(1-p)p}} \leq K$. Thus,

$$\sqrt{\frac{1-p}{np}} \leq K \leq (C^* \log^2 n)^{-1}. \tag{19}$$

Then, K exists if $\sqrt{\dfrac{1-p}{np}} \leq (C^* \log^2 n)^{-1}$ or if p satisfies the more stringent constraint

$$p \geq C' n^{-1} \log^4 n,$$

where C' is a constant depending on C^*. Inequality (18) is obtained from (17) setting $K = \sqrt{\dfrac{1-p}{np}}$. ■

Let us observe that for spectral norm of matrix $\widetilde{A}^{\mathrm{ER}}$ is lower bounded by the extreme of the support of the limiting e.s.d. $F^{\widetilde{A}^{\mathrm{ER}}}(x)$. Additionally, if for any $\delta > 0$, $p > \delta n^{-1} \log^4 n$, then the spectral norm is concentrated around the extreme of the support of $F^{\widetilde{A}^{\mathrm{ER}}}(x)$.

4.4 Spectrum of the Normalized ER Adjacency Matrix

In the previous Section (4.2) and (4.3) we focused on the spectral properties of the normalized centered ER adjacency matrix $\widetilde{A}^{\mathrm{ER}}$ while in this section we analyze the spectral properties of the normalized ER adjacency matrix $\widehat{A}^{\mathrm{ER}}$ and the effect of the mean component $\overline{A}^{\mathrm{ER}}$ on it. The following lemma plays a key role to establish a fundamental relation between the eigenvalue e.d.f. $F^{\widetilde{A}^{\mathrm{ER}}}(x)$ studied in the previous sections and $F^{\widehat{A}^{\mathrm{ER}}}(x)$.

LEMMA 2 *[3] If $F^A(x)$, $F^B(x)$ are the eigenvalue e.d.f. of A , B, Hermitian matrices of size n, then*

$$|F^A(x) - F^B(x)| \leq \frac{\mathrm{rank}(A-B)}{n}.$$

We recall that $\overline{A}^{\mathrm{ER}} = \widehat{A}^{\mathrm{ER}} - \widetilde{A}^{\mathrm{ER}}$ has unit rank for any n. Then, asymptotically for $n \to \infty$, the limiting eigenvalue distribution of the matrix $\widehat{A}^{\mathrm{ER}}$ converges to the semicircular law as the limiting eigenvalue distribution of the matrix $\widetilde{A}^{\mathrm{ER}}$. As for $\widetilde{A}^{\mathrm{ER}}$, the limiting eigenvalue distribution of the matrix $\widehat{A}^{\mathrm{ER}}$ provides only a lower bound for the spectral norm that requires independent study.

The spectral norm of the two matrices $\widehat{A}^{\mathrm{ER}}$ and $\widetilde{A}^{\mathrm{ER}}$ are different, because the largest eigenvalue changes when a unit rank matrix is added to a Hermitian matrix. From Bauer-Fike theorem for Hermitian matrices [19], we have

$$|\lambda_i(\widehat{A}^{\mathrm{ER}}) - \lambda_i(\overline{A}^{\mathrm{ER}})| \leq ||\widetilde{A}^{\mathrm{ER}}||_2 \tag{20}$$

for $1 \leq i \leq n$.

Let us note that $\lambda_1(\overline{A}^{\mathrm{ER}}) = n\gamma(n)p_n$ and $\lambda_i(\overline{A}^{\mathrm{ER}}) = 0$ for $i \geq 2$. Then, from (20), asymptotically as $n \to +\infty$, $\|\widetilde{A}^{\mathrm{ER}}\|_2 \overset{a.s.}{\to} 2$. Thus, we get the following concentration result for the largest eigenvalue of the perturbed matrix $\widehat{A}^{\mathrm{ER}}$

$$\left| \lambda_1(\widehat{A}^{\mathrm{ER}}) - \sqrt{\frac{np(n)}{1-p(n)}} \right| \leq 2. \tag{21}$$

We notice that, for dense and sparse networks, $\sqrt{\frac{np(n)}{1-p(n)}} \gg 2$. Hence the above result implies that $\lambda_1(\widehat{A}^{\mathrm{ER}}) \to n\gamma(n)p_n$. Thus, the following lemma holds.

Lemma 3 *[11] Let $\widehat{A}^{ER}$ satisfies the conditions of Corollary 2. Then,*

$$\lambda_1^n(\widehat{A}^{\mathrm{ER}}) \xrightarrow{n\to\infty, a.s.} \sqrt{\frac{np(n)}{1-p(n)}},$$

i.e., the largest eigenvalue of $\widehat{A}^{\mathrm{ER}}$ tends to the largest eigenvalue of the mean matrix $\overline{A}^{\mathrm{ER}}$ as $n \to \infty$.

4.5 Limiting Spectral Distribution of Centered Hermitian Matrices

In this section we present a useful existing result on centered symmetric matrices H with independent upper diagonal entries whose distributions have in general different variances. This result provides the Stieltjes transform of the limiting eigenvalue distribution as $n \to +\infty$ as solution of a nonlinear system of n equation.

Proposition 2 *[15, 16, Chapter 1] Let the symmetric matrix H satisfy Assumption 1. Additionally, the variances σ_{ij}^2 of its entries satisfy the conditions*

$$\sup_n \max_{i=1,2,..n} \sum_j \sigma_{ij}^2 < \infty$$

and $\inf_{i,j} n\sigma_{ij}^2 = c > 0$. Then, as $n \to +\infty$, almost surely $F^H(x,n)$, the random e.s.d. of the $n \times n$ matrix H converges for any x to a deterministic distribution function $S_n^H(x)$ whose Stieltjes transform $s(z)$ is given by

$$s_n(z) = \int \frac{\mathrm{d}S_n^H(x)}{x-z} = \frac{1}{n}\sum_{i=1}^n c_i(z,n),$$

where $c_i(z,n)$ is the unique solution to the system of equations

$$c_i(z,n) = \left\{ \left[-zI - \left(\delta_{pl} \sum_{s=1}^n c_s(z,n)\sigma_{sl}^2 \right)_{p,l=1}^n \right]^{-1} \right\}_{ii}, \qquad i = 1, 2, ..n$$

in the class of analytic functions

$$\mathcal{A} = \{\Im(z)\Im(c_i(z,n)) > 0, \Im(z) \neq 0\}.$$

5 Results for Adjacency Matrix of M community Model

5.1 Finding the Spectrum of Centered Adjacency Matrix

In this section we consider the normalized centered SBM adjacency matrix $\widetilde{A}$ with $\gamma(n) = (np^*(1-p^*))^{-1}$ where $p^* = \max_{m=1,\dots M} p_m$. Additionally, we assume that all the probabilities p_m scales at the same rate, i.e. $\lim_{n\to+\infty} \frac{p_i}{p_j} = c_{ij}$ for some $c_{ij} > 0$, and $p_m(n) \in \omega(n^{-1})$ as $n \to +\infty$. Then, it is straightforward to verify that the conditions of Proposition 2 are satisfied and the following corollary holds.

COROLLARY 3 *Let $\widetilde{A}$ be the normalized centered SBM adjacency matrix defined in (8) with $\gamma(n) = (np^*(1-p^*))^{-1}$ and $p^* = \max_{m=1,\dots M} p_m$. If $p_m(n) \in \omega(n^{-1})$ and $p_m(n) = O(p_0(n))$ for all $m = 1, \dots M$, then, almost surely, the eigenvalue e.d.f. converge weakly to a distribution function whose Stieltjes transform is given by*

$$s(z) = \sum_{m=1}^{M} c_m(z) \tag{22}$$

$c_m(z)$ being the unique solution to the system of equation

$$c_m(z) = \frac{-1/M}{z + \varsigma_m c_m(z) + \varsigma_0 \sum_{\ell \neq m} c_\ell(z)}, \tag{23}$$

with $\varsigma_\ell = \lim_{n\to+\infty} \dfrac{p_\ell(1-p_\ell)}{p^*(1-p^*)}$, that satisfies the conditions

$$\Im(c_\ell(z))\Im(z) > 0 \text{ for } \Im z > 0. \tag{24}$$

The above result implies that in general the asymptotic eigenvalue e.d.f. of an SBM is not a semicircular law any longer.

5.2 Spectrum of the Full Adjacency Matrix

The result above gives the spectrum of the matrix $\widetilde{A}$. Let us recall the definition of matrix $\widehat{A}$ as given in (5)

$$\widehat{A} = \widetilde{A} + \overline{A}.$$

Using Lemma 2 on the finite rank perturbation, we deduce that the asymptotic eigenvalue e.d.f. of $\widehat{A}$ and $\widetilde{A}$ are the same. However, their spectra differ in the extreme eigenvalues.

5.3 Extreme Eigenvalue of Adjacency Matrix

Using the Bauer-Fike Theorem on perturbation of eigenvalues of Hermitian matrices, for matrices $\widehat{A}$, $\overline{A}$, and $\widetilde{A}$, we have

$$|\lambda_i(\widehat{A}) - \lambda_i(\overline{A})| \leq \|\widetilde{A}\|_2.$$

This is useful in getting asymptotic characterization of the M largest eigenvalues of $\widehat{A}$. Since $\overline{A}$ has M non-zero eigenvalues, this result says that the first M largest eigenvalues of $\widehat{A}$ are concentrated around these eigenvalues, within an error of the order of the spectral norm of $\widetilde{A}$. The other eigenvalues of $\widehat{A}$ are below the spectral norm of $\widetilde{A}$, and hence they fall inside the continous part of the limiting e.d.f. To use this result, we need a bound on the spectral norm of the centered SBM adjacency matrix $\widetilde{A}$. We strengthen the result in Theorem 1.4 in [21] to derive a tighter bound on the spectral norm of $\widetilde{A}$.

THEOREM 1 *Let $\widetilde{A}$ be a normalized centered SBM adjacency matrix defined in (9) and satisfying the same conditions as in Corollary 3. Additionally, $p_0(n)$ satisfies the inequality $p_0(n) \geq C'n^{-1}\log^4 n$ for some constant $C' > 0$ and $\sup \frac{p_0(n)(1-p_0(n))}{p^*(n)(1-p^*(n))}$ is bounded. Then, there exists a constant $C > 0$ such that almost surely*

$$\|\widetilde{A}\|_2 \leq 2\sqrt{M^{-1}\left(1+(M-1)\varsigma_0\right)} + C\sqrt[4]{\frac{1-p_0(n)}{np_0(n)}}\log(n)$$

with ς_0 defined as in Corollary 3.

Proof: Much of the proof follows that of Theorem 1.3 in [21]. We first bound the spectral norm of the unnormalized centered adjacency matrix $\widetilde{A}' = \widetilde{A}/\gamma(n)$. At the end of the proof, we use this bound to derive a bound on the spectral norm of $\widetilde{A}$, which gives us the result. We use the idea that spectral norm, which is the largest dominating eigenvalue in absolute value, can be bounded by the trace of the matrix raised to an even exponent, and that the larger the exponent, the sharper the bound:

$$||\widetilde{A}'||_2^{2k} = \max_{1\leq i\leq n} |\lambda_i(\widetilde{A}')|^{2k} \leq (\sum_{i=1}^{n} |\lambda_i(\widetilde{A}')|^{2k}) = (\mathrm{tr}(\widetilde{A}')^{2k}).$$

The method of proof is the popular moment method orginated by Füredi and Komlós, which bounds the moments of the e.s.d. of the matrix by bounding the expected trace of the matrix using combinatorial arguments [14].

Once we obtain a bound on the expected spectral norm, we can use Markov inequality, to bound the tail probabilities.

$$\Pr\{||\widetilde{A}'||_2 \geq \lambda\} \leq \frac{E||\widetilde{A}'||_2^{2k}}{\lambda^{2k}} \leq \frac{E\mathrm{tr}(\widetilde{A}')^{2k}}{\lambda^{2k}} \tag{25}$$

If A is a standard Wigner matrix, for fixed k, the right hand side of the above equation is n times the $2k^{\text{th}}$ moment of the empirical spectral distribution, which by Semicircle law tends to C_k. Therefore, for such matrices, if k were chosen to be a fixed number independent of n, the right hand side tends to infinity for large n, making it rather useless. Therefore we choose k to be a function of n. The idea is that when k is a slowly increasing function of n, the semicircle law still holds, and since $C_k \leq 4^k$, the upper bound tends to 0, for any $\lambda \geq 2$. Here, we extend this idea to Wigner matrix displaying community structure.

Next, we need to find a bound on the quantity $E\mathrm{tr}(\widetilde{A}')^{2k}$. To do this we expand the trace as a summation of expectation over cycles of length $2k$ of vertices in the set $\{1, 2, 3, ...n\}$

$$E\mathrm{tr}(\widetilde{A}')^{2k} = E\sum_{i_1,i_2,i_3,....i_{2k}} X_{i_1,i_2}X_{i_2,i_3}....X_{i_{2k},i_1},$$

where $\{i_1, i_2...i_{2k}, i_1\}$ form a cycle over edges such that $i_j \in \{1, 2, 3, ...n\}$, for each $1 \leq j \leq 2k$. Each edge $\{i_{j-1}, i_j\}$ corresponds to a random variable X_{i_{j-1},i_j}.

We can partition the graphs based on the number of unique vertices that appear in the graph, called the weight of the graph, denoted by t, $1 \leq t \leq 2k$. We can represent the original graph on $2k$ edges and $2k$ vertices equivalently by using a condensed undirected connected graph on t vertices. An edge exists in this graph if and only if it exists in the original graph and if it exists more than once in the original graph(walk), then this edge has a weight equal to the number of times this edge is traversed by the walk. Since each such random variable is zero mean and since

each variable is independent, if an undirected edge $\{i_{j-1}, i_j\}$ has a weight equal to 1, that is it only appears once in the walk, then the whole term becomes zero. So we need only consider the contribution of those graphs that have every edge appearing at least twice.

By virtue of independence and zero mean property, if t is greater than $k+1$, and because the number of edges required for connectivity is greater than or equal to $t+1$, there must at least be $k+1$ edges in the graph. Since the total number of edges is $2k$ in the walk, this means there exists an edge that appears only once, making the contribution of such a term zero. Hence we must have $1 \leq t \leq k+1$. Thus we can bound the quantity of interest as below:

$$E\mathrm{tr}(\widetilde{A}')^{2k} \leq \sum_{t=1}^{k+1} \sum_{G \in \mathcal{G}_{t,n,2k}} EX_G,$$

where $\mathcal{G}_{t,n}$ represents a set of graphs on t vertices drawn from $\{1, 2, ..n\}$ with $2k$ edges.

An edge $e = \{i_{j-1}, i_j\}$ such as described above, is called an innovation edge, if the vertex i_j is such that $i_j \notin \{i_1, i_2, ...i_{j-1}\}$, i.e., it appears for the first time in e. The other edges in the graph either overlap the innovation edges or are interconnections between two vertices that already exist in the graph. Since in our case the random variables are bounded in absolute value by 1(since they are Bernoulli), the contribution over all the edges other than the innovation edges can be bounded by 1. For any graph on t vertices, there must be exactly $t-1$ innovation edges and each edge must have at least weight 2. The contribution to the expectation of each such edge would have a weight of at most σ_i^2, $0 \leq i \leq M$ depending on whether the edge is between two communities or within some community i. This bound is exact if each such edge has a weight two; if it has weight more than two, then this is an upper bound. Then, by independence, the contribution of the group of edges can be bounded by the product. Therefore, the following is true:

$$|Ea_{i1,i2}a_{i2,i3}...a_{i_{2k},i_1}| \leq (\max_{1 \leq i \leq M} \sigma_i^2)^{(t-1-i)}(\sigma_0^2)^i,$$

where i are integers such that $0 \leq i \leq t-1$.

This corresponds to choosing out of t edges, i edges such that the vertices of those edges belong to two different communities. Once we choose such i edges we need to choose the communities from which the corresponding vertices emerge. For convenience, we can assume the vertices have a preferred ordering. The first such vertex can then be chosen from any of the M communities. Once such a community is chosen, if the next edge is upper bounded by $\max_i \sigma_i^2$, the next vertex of the edge can be chosen only in 1 way, because this corresponds to an edge belonging to the same community. If the edge is bounded by σ_o^2, then the community to which the next vertex belongs can be chosen in at most $M-1$ ways, since this edge corresponds to an edge between communities. Corresponding to each selection of a community to which the edge can belong, the vertices can be chosen in $(n/M)^t$ ways.

Therefore, we can finally bound the full term as follows:

$$E\mathrm{tr}\widetilde{A}'^{2k} \leq \sum_{t=1}^{k+1} M(n/M)^t \sum_{i=0}^{t-1} \binom{t-1}{i} (M-1)^{(t-1-i)}(\sigma_0^2)^{(t-1-i)}(\max_{1 \leq j \leq m} \sigma_j^2)^i W(k,t).$$

The inner summation on the variable i turns out to be the Binomial expansion. $W(k,t)$ is the number of equivalent graphs of t fixed vertices with $2k$ edges and is related to the number of the Catalan number C_t. We use a bound on this quantity which is available in [21]:

$$W(k,t) \leq \binom{2k}{2p-2} p^N 2^{k+N+1}(N+2)^N,$$

where $N = 2k - 2(t-1)$. Finally, we get:

$$E\mathrm{tr}\widetilde{A}'^{2k} \leq \sum_{t=0}^{k+1} n^t (\sigma^2 n)^{t-1} W(k,t),$$

where,

$$\sigma^2 = \frac{1}{M}(\max_i \sigma_i^2 + (M-1)\sigma_o^2). \tag{26}$$

As in [21] it can be shown that when $2k = a\sigma^{1/2}n^{1/4}$, for some a, the term within the summation is bounded by a geometric series with growth factor $1/2$. Using this fact we finally obtain:

$$E\mathrm{tr}(\widetilde{A}')^{2k} \leq 2n(2\sigma\sqrt{n})^{2k}.$$

Substituting the above in equation (25), and using $\lambda = 2\sigma\sqrt{n} + C(\sigma)^{1/2}n^{1/4}\log(n)$ we have:

$$\begin{aligned}
\Pr\{||\widetilde{A}'||_2 \geq 2\sigma\sqrt{n} + C(\sigma)^{1/2}n^{1/4}\log(n)\} &\leq 2n\left(\frac{2\sigma\sqrt{n}}{2\sigma\sqrt{n} + C(\sigma)^{1/2}n^{1/4}\log(n)}\right)^{2k} \\
&= 2n(1 - \frac{C\sigma^{1/2}n^{1/4}\log(n)}{2\sigma\sqrt{n} + C\sigma^{1/2}n^{1/4}\log(n)})^{2k} \\
&\leq 2n(1 - \frac{C\sigma^{1/2}n^{1/4}\log(n)}{3\sigma\sqrt{n}})^{2k} \\
&\leq 2n\exp(-\frac{c\log(n)k}{3\sqrt{\sigma}n^{1/4}}) \\
&= 2n\exp(-ca\log(n)/3),
\end{aligned}$$

where the second inequality above follows from the assumption that $\sigma \geq C'n^{-1/2}\log^2(n)$, the third inequality because $e^{-x} \geq 1 - x$, and the last equality by the form of k. Now since the right hand side is summable in n for appropriate constants c and a, by Borel-Cantelli Lemma [4], we have:

$$||\widetilde{A}'|| \leq 2\sigma\sqrt{n} + C(\sigma)^{1/2}n^{1/4}\log(n) \text{ a.s.}$$

for $\sigma \geq C'Kn^{-1/2}\log^2(n)$, where $K = 1$. This is the same identity as in [21] except with the appropriate definition of σ from (26).

In the above, since $\widetilde{A}'$ is unnormalized, we were able to upper bound each variable by $K = 1$. It is easy to redo the above proof for the matrix $\widetilde{A}$, where because of the normalization, we have the upper bound on the matrix entries, $K = \frac{1-p_0}{\sqrt{np^*(1-p^*)}}$. One can easily see that the condition $\sigma \geq KC'n^{-1/2}\log^2(n)$ holds when $p_0 \geq C'n^{-1}\log^4(n)$, and the bound on $\widetilde{A}$ becomes:

$$\begin{aligned}
||\widetilde{A}||_2 &\leq 2\sigma\sqrt{n} + C(K\sigma)^{1/2}n^{1/4}\log(n) \\
&\leq 2\frac{1}{M}(1 + (M-1)\varsigma_0) + C\left(\frac{1-p_0}{\sqrt{np_0(1-p_0)}}\right)^{1/2}\left(\frac{p_0(1-p_0)}{p^*(1-p^*)}\right)^{1/4} \\
&\quad \left(\frac{(M-1)p_0(1-p_0) + p^*(1-p^*)}{np^*(1-p^*)}\right)^{1/4} n^{1/4}\log(n) \\
&\leq 2\frac{1}{M}(1 + (M-1)\varsigma_0) + \frac{C}{n^{1/4}}\left(\frac{1-p_0}{p_0}\right)^{1/4}\log(n),
\end{aligned}$$

where we used the fact that $\sup_n \frac{p_0(1-p_0)}{p^*(1-p^*)}$ is finite, and C is an arbitrary constant. ■

5.4 Eigenvalues of the Mean Matrix

By the result above on the spectral norm of the zero mean matrix, we know that the largest eigenvalue of the matrix is somewhere close to the edge of the spectrum. But when the mean matrix is added to this matrix, the largest eigenvalue escapes the bounded spectrum. Namely, since the mean matrix has rank M, by interlacing inequalities on the sum of two Hermitian matrices, we can see that there are exactly M eigenvalues outside the bounded spectrum. Recall that

$$\widehat{A} = \widetilde{A} + \overline{A}$$

By the Bauer-Fike Theorem we have

$$|\lambda_i(\widehat{A}) - \lambda_i(\overline{A})| \leq \|\widetilde{A}\|. \tag{27}$$

From Theorem 1 we have that asymptotically almost surely $\|\widetilde{A}\| \leq 2\varsigma + \delta$ with $\varsigma = \sqrt{M^{-1}(1 + (M-1)\varsigma_0)}$ and $\delta \to 0$. For $i > M$, $\lambda_i(\overline{A}) = 0$. Therefore, we see that $\lambda_i(\widehat{A})$ for $i > M$ lies below the spectral norm of $\widetilde{A}$.

5.4.1 Eigenvalues of $\overline{A}$

Let the eigenvalues of P in (6) be $\mu_i, 1 \leq i \leq M$. They depend on the probabilities p_i, $0 \leq i \leq M$ and the following relationship holds between the eigenvalues of $\widehat{A}$ and μ_i.

LEMMA 4 *Under the conditions in Proposition 1 the M eigenvalues of $\widehat{A}$, outside the continuous spectrum of $\widehat{A}$ are given as:*

$$|\lambda_i(\widehat{A}) - \mu_i| \leq 2\varsigma + \delta \tag{28}$$

for $1 \leq i \leq M$.

Thus, we conclude that asymptotically, the M largest eigenvalues of the adjacency matrix converge to those of the mean matrix almost surely.

To complete this argument, we need the approximate locations of μ_i's. By Gershgorin disc theorem [19], the μ_i's should satisfy:

$$\left| \frac{\mu_i M}{\gamma(n) n} - p_i \right| \leq p_0(M-1). \tag{29}$$

Note on special case of Symmetric SBM:

When the probabilities $p_1 = p_2 = p_3 \ldots = p_M = p^*$, we can significantly simplify the equations (23) and achieve useful insight into the shape of the spectrum. In this case we have that $\varsigma_m = 1$ for $m = 1, 2, ..M$ and the fixed point equation (23) becomes:

$$Mc_m(z) = \frac{-1}{z + c_m(z) + \varsigma_0 \sum_{\ell \neq m} c_\ell(z)}.$$

We see that the equations are symmetric, hence, by uniqueness property of Corollary 3, we must have that $c_1(z) = c_2(z) = c_3(z) = \ldots = c_M(z) = c(z)$, and $s(z) = Mc(z)$, which leads to

$$s(z) = \frac{-1}{z + \frac{(1+(M-1)\varsigma_0)}{M} s(z)}.$$

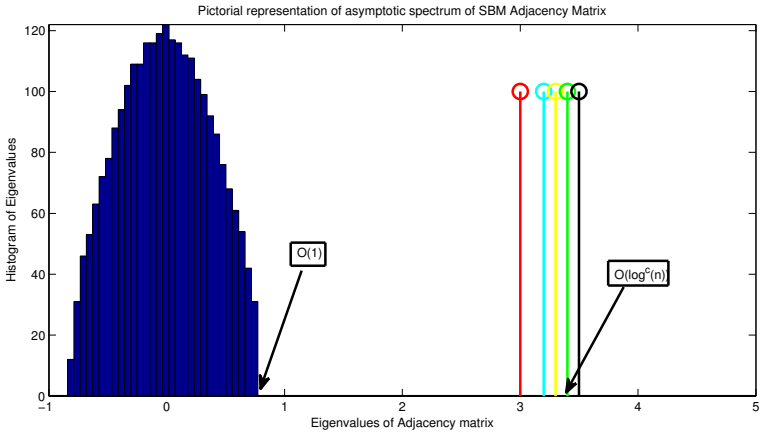


Figure 1: Plot showing approximate asymptotic eigenvalue locations of the adjacency Matrix

This is the same as the equation for the Stieltjes transform of the semicircular law given in equation (15), with $\sigma^2 \equiv \frac{(1+(M-1)\varsigma_0)}{M}$. Thus, we see that in the symmetric scenario, the spectrum of the adjacency matrix becomes a semicircle law, and the upper bound in Theorem 1 becomes exact. Similarly, the eigenvalues of the mean matrix become:

$$\mu_1 = \frac{\sqrt{n}}{M\sqrt{p_1(1-p_1)}}(p_1 + (M-1)p_0). \tag{30}$$

and

$$\mu_i = \frac{\sqrt{n}}{M\sqrt{p_1(1-p_1)}}(p_1 - p_0). \tag{31}$$

for $i = 2, 3, \ldots, M$. Thus, from Lemma 4, the largest eigenvalue of the $\widehat{A}$ converges to (30) above, and the next $M-1$ largest eigenvalues converge to (31), i.e., the second largest eigenvalue of the adjacency matrix has multiplicity $M-1$.

In Figure 1 we show diagramatically the general form of the asymptotic histogram of the scaled adjacency matrix of an SBM, with approximate locations of its various components all probabilities scale as $\log^c(n)$ for some $c > 4$. We observe that under these conditions, there is sufficient separation between the continuous part of the spectrum and the discrete extremal eigenvalues.

6 Spectral Distribution of Normalized Laplacian Matrix

We recall that the normalized Laplacian Matrix is given by

$$\mathcal{L} = I - D^{-1/2}AD^{-1/2}. \tag{32}$$

For the sake of simplicity, we consider the case of two blocks, i.e., $M = 2$, and probabilities $p_i, 0 \leq i \leq 2$ that are not dependent on the size of the matrix n.

Let $P^{'} = D^{-1/2}AD^{-1/2}$. We show that asymptotically, the e.s.d. of the matrix $\frac{1}{2}\sqrt{n}P^{'}$ converges to the e.s.d. of the matrix $\frac{1}{\sqrt{n}}A^{''}$, defined as

$$A^{''}_{ij} = \begin{cases} A_{ij}/(p_1+p_0), & \text{if } i,j \in \Omega_1 \\ A_{ij}/(p_2+p_0), & \text{if } i,j \in \Omega_2 \\ A_{ij}/\sqrt{(p_1+p_0)(p_2+p_0)} & \text{otherwise} \end{cases}$$

Consequently, the following holds.

LEMMA 5 *The distribution of matrix* $\frac{1}{2}\sqrt{n}P^{'}$ *is given by:*

$$c_i = \frac{-1/2}{z + \sigma_i^{'} c_i + \sigma_0^{'} c_j}, \tag{33}$$

for $i, j = 1, 2$ *and* $i, j = 2, 1$ *respectively. where* $\sigma_1^{'} = \frac{\sigma_1^2}{(p_1+p_0)^2}$, $\sigma_2^{'} = \frac{\sigma_2^2}{(p_2+p_0)^2}$, $\sigma_0^{'} = \frac{\sigma_0^2}{(p_0+p_1)(p_0+p_2)}$ *and the limiting distribution has a spectrum whose Stieltjes transform is given by* $c(z) = c_1(z) + c_2(z)$.

Since $\frac{\sqrt{n}}{2}\mathcal{L} = \frac{\sqrt{n}}{2} - \frac{\sqrt{n}}{2}P^{'}$, its distribution has a bulk component that lies around $\sqrt{n}/2$, with an approximate width of $2\sqrt{\max(\sigma_1^{'}, \sigma_2^{'}) + \sigma_0^{'}}$. This matrix also has an eigenvalue at 0, by the property of Laplacian.

In the two-community case, it can be seen from simulations that there exists one more eigenvalue outside the bulk, which remains to be properly characterized.

Proof of the lemma: We follow the steps in the proof of Theorem 1.1 in [5]. The first step is a form of uniform strong law of large numbers called Kolomogorov-Marcinkiewicz-Zygmund strong law of large numbers. Since the elements of the matrix A are independent and have finite fourth moments from Lemma 2.3 in [5] we have the following as true:

$$\sum_{j=1}^{n} A_{ij} = \frac{n}{2}(p_1 + p_0 + \delta_i^{(1)}), \tag{34}$$

where $\max_i |\delta_i^1| = o(1)$ for $1 \leq i \leq n/2$ and

$$\sum_{j=1}^{n} A_{ij} = \frac{n}{2}(p_2 + p_0 + \delta_i^{(2)}), \tag{35}$$

where $\max_i |\delta_i^2| = o(1)$ for $1 + n/2 \leq i \leq n$.

From equations (34) and (35) we have uniform convergence for $1 \leq i \leq n$, with the error, $\max_i(\delta_i^{(1)}, \delta_i^{(2)}) = o(1)$:

$$D_i = \sum_{j=1}^{n} A_{ij} = \frac{n}{2}(p_k + p_0) + \epsilon_i, \tag{36}$$

where $p_k = p_1$ if $i \in \Omega_1$ and $p_k = p_2$ if $i \in \Omega_2$, and $\max_i |\epsilon_i| = \epsilon = o(1)$ uniformly.

Next step is to use Hoffman-Wielandt inequality [1] to bound the error between the e.s.d. of $\frac{A^{''}}{\sqrt{n}}$ and $\frac{\sqrt{n}}{2}P^{'}$.

We have using Hoffman-Wielandt inequality and bound on Stieltjes transforms found in [2]:

$$|s_{F^{\frac{1}{\sqrt{n}}A^{''}}}(z) - s_{F^{\frac{\sqrt{n}}{2}P^{'}}}(z)| \leq \frac{c}{n\Im z} \sum_{ij} |\frac{\sqrt{n}}{2}P_{ij}^{'} - \frac{1}{\sqrt{n}}A_{ij}^{''}|^2,$$

where $z \in \mathbb{C}^+$.

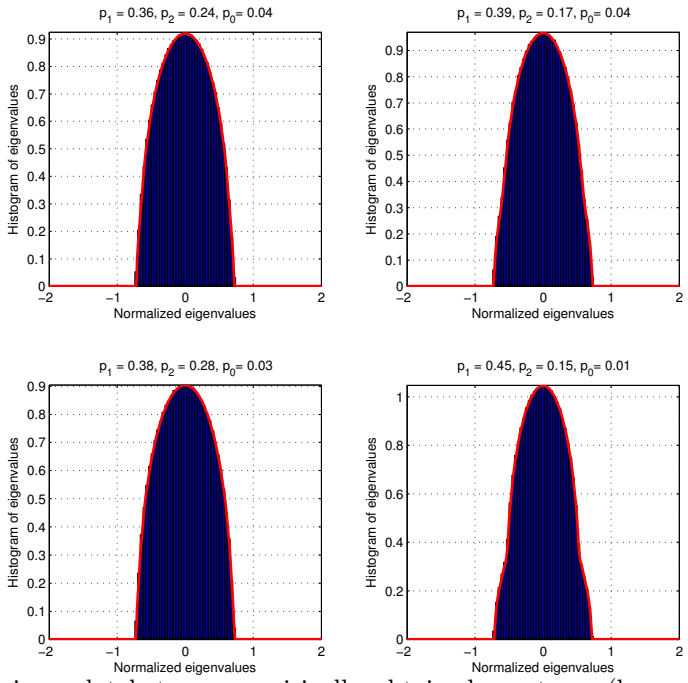


Figure 2: Comparison plot between empirically obtained spectrum (bar graph), and explicit solution(line) of 2-community SBM adjacency matrix

For any i, j, we have

$$\begin{aligned}\frac{\sqrt{n}}{2}P'_{ij} &= \frac{\sqrt{n}}{2}\frac{A_{ij}}{\sqrt{D_i D_j}} \\ &= \frac{\sqrt{n}}{2}\frac{A_{ij}}{\sqrt{n/2(p_k+p_0+\epsilon_i)n/2(p_l+p_0+\epsilon_j)}} \\ &= \frac{A_{ij}}{\sqrt{n(p_k+p_0)(p_l+p_0)}}(1+O(\epsilon_i))(1+O(\epsilon_j)) \\ &= \frac{A''_{ij}}{\sqrt{n}}(1+O(\epsilon)), \end{aligned} \tag{37}$$

where $i \in \Omega_k$ and $j \in \Omega_l$ and ϵ is infinitesimally small. The last equality follows because ϵ_i and ϵ_j tend to 0 uniformly for large n. Thus by Hoffman Wielandt inequality we have:

$$|s_{F^{\frac{1}{\sqrt{n}}A''}}(z) - s_{F^{\frac{\sqrt{n}}{2}P'}}(z)| \leq \frac{c}{n^2 \Im z}\sum_{ij}|A''_{ij}|^2 O(\epsilon^2) \to 0$$

a.s.

The last relation follows from the strong law of large numbers on variables A''_{ij} and since $\epsilon = o(1)$. Hence we have the result. ■

7 Example Application: Epidemic Spreading

In this section, we discuss an important potential application of the result we derived above for adjacency matrices, namely, in the topic of epidemic spreading. We refer to the recent paper [7].

In this work, the authors study an epidemic process over a random network of nodes. The spread of the epidemic from one node to another is governed by the Random network, i.e., a node can only infect another if there exists an edge between the two nodes. They present a concise result delineating the relationship between the expected cost of the epidemic per node denoted by $C_D(n)$(disease cost) [7], and the largest eigenvalue of the modified adjacency matrix; namely,

$$C_D(n) \leq \frac{\alpha c_d}{1 - \lambda_1(M)}, \tag{38}$$

where $M = (1 - \delta)I + \beta A$ is the matrix which governs the dynamics of the system [7], with β being the probability of infection, δ is the probability of recovery of any node, and c_d is the cost parameter. We direct the reader to the original paper for more details. A is as usual the adjacency matrix of the random graph.
We examine the epidemic spread on a graph which follows SBM with M communities. We know that in this case $\lambda_1(A) \to n/M\mu_1$ as $n \to \infty$ a.s. under certain conditions. Also by (29) we have that $\mu_1 \leq p_1 + (M-1)p_0$, therefore we obtain:

$$\lambda_1(M) = (1 - \delta) + \beta\lambda_1(A) \leq 1 - \delta + \beta(n/M\mu_1).$$

This yields,

$$C_D(n) \leq \frac{\alpha c_d}{\delta - \beta n/M(p_1 + (M-1)p_0)}. \tag{39}$$

If $p_1 \gg p_i$, for $i \geq 2$, then we can venture to say that this bound is tight, and that the community with the largest edge probability governs the disease cost.

8 Numerical Results

In this section we provide simulation results to demonstrate the results obtained above. More specifically, we corroborate our results on the spectrum of adjacency matrix by comparing the spectrum obtained by simulating a 2-community SBM with the distribution obtained by inverting the Stieltjes transform, which is an explicit solution of the simultaneous equations (23). In the simulations, we use a matrix of size $n = 10^4$. For a 2-community system, the solution amounts to solving explicitly the resulting quartic equation and choosing the solution branch that satisfies the conditions (24). The inverse relationship between the limiting e.s.d. and the Stieltjes transform thus obtained, is given by the well known Stieltjes inversion formula:

$$f(x) = \lim_{y \to 0} \Im s_F(x + \sqrt{-1}y)/\pi, \tag{40}$$

where $f(x)$ is the p.d.f. corresponding to the c.d.f. $F(x)$, whenever the limit exists.
Figure 2 shows the histogram of normalized adjacency matrix $\frac{1}{\sqrt{n}}A$ and compares it to the theoretical spectrum obtained as above for $n = 10^4$, and several values of edge probabilities.

In the second part of this section we turn our attention to the extremal eigenvalues of the adjacency matrix for a 3-community SBM of size $n = 999$. Over several independent runs, we get values of the top 4 eigenvalues of the matrix A, for $0.3 \leq p_1 \leq 0.48$, $0.15 \leq p_2 \leq 0.33$, $0.08 \leq p_3 \leq 0.26$ and $0.03 \leq p_0 \leq 0.031$, randomly picked. We note that as expected in (Figure 3), there are three eigenvalues outside the bulk, which agree very well with the expected values, i.e., the non-zero eigenvalues of $\overline{A}$. In addition, it can also be seen that the upper bound in Theorem 1 is remarkably tight for the simulated probabilities.

Next, we consider the spectrum of the normalized Laplacian matrix. In fact we consider the spectrum of the following matrix which we denote $\tilde{\mathcal{L}}$, which is given by:$\tilde{\mathcal{L}} = \sqrt{n}/2 - \sqrt{n}/2\mathcal{L}$. By

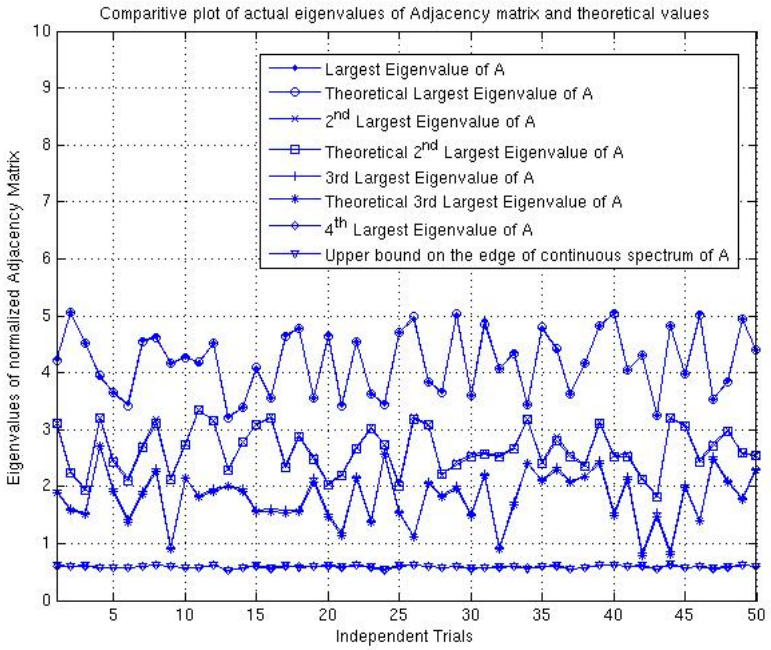


Figure 3: Extremal eigenvalues of 3-community SBM normalized matrix compared to expected values.

assertion its spectrum is given by the solution of (33). We explicitly solve this equation for a two-community model, and compare it to numerically obtained results for a matrix whose size is $n = 999$ for various values of the probabilities p_1, p_2 and p_0 (Fig.4).

9 Conclusion

In this work we studied in detail the spectra of adjacency and normalized Laplacian matrices of an SBM with M communities. In particular, we analyzed the limiting empirical distribution of the eigenvalues of the adjacency matrix of SBM. We find that the Stieltjes transform of the limiting distribution satisfies a fixed point equation and provide an explicit expression in the case of symmetric communities. Furthermore, we obtained a tight bound on the support of the asymptotic spectrum, and concentration bounds on the extremal eigenvalues.

As future work we plan to analyze the structure of the eigenvectors and develop more detailed applications to graph clustering and network sampling. It will be also interesting to consider SBM models where sizes of communities are not uniform.

10 Acknowledgement

This work was partly funded by the French Government (National Research Agency, ANR) through the “Investments for the Future” Program reference #ANR-11-LABX-0031-01.

Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme «Investissements d'Avenir »portant la référence : ANR-11-LABX-0031-01.

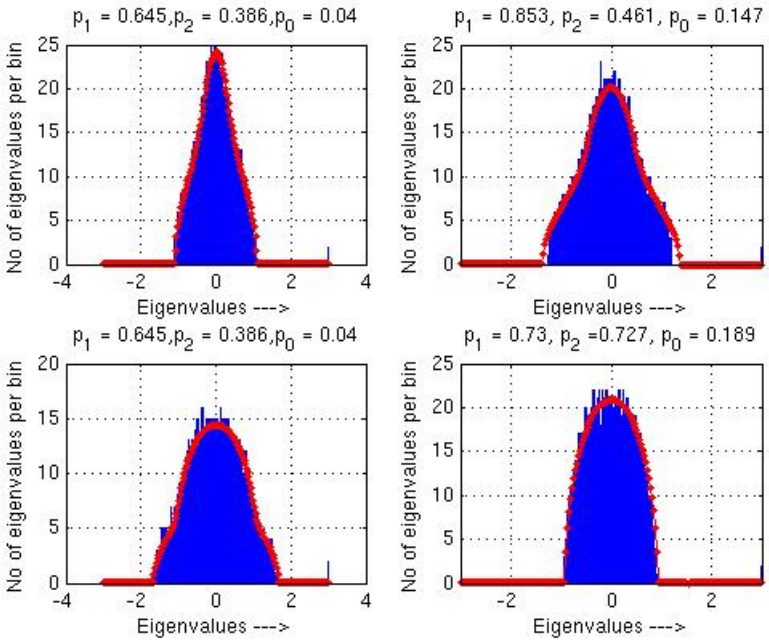


Figure 4: Histogram of 2-community $\tilde{\mathcal{L}}$ for various edge probabilities compared to theoretical spectrum

References

[1] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*. Cambridge University Press, 2009.

[2] Z. D. Bai, "Methodologies in spectral analysis of large dimensional random matrices, a review," *Statistica Sinica*, vol. 9, no. 3, pp. 611–677, Jul. 1999.

[3] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*. Springer, 2009.

[4] P. Billingsley, *Probability and Measure*, 3rd ed. New York, NY: Wiley, 1995.

[5] C. Bordanave, P. Caputo, and D. Chafaï, "Spectrum of large random reversible markov chains: Two examples," *Latin American Journal of Probability and Mathematical Statistics*, vol. 7, pp. 41–64, 2010.

[6] C. Bordenave and M. Lelarge, "Resolvent of large random graphs," *Random Structures and Algorithms*, vol. 37, no. 3, pp. 332–352, October 2010.

[7] S. Bose, E. Bodine-Baron, B. Hassibi, and A. Wierman, "The cost of an epidemic over a complex network: A random matrix approach," *Mathematics of Operations Research*, 2014.

[8] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[9] A. Condon and R. Karp, "Algorithms for graph partitioning on the planted partition model," *Random Structures and Algorithms*, vol. 18, pp. 116–140, 2001.

[10] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Physical Review E*, vol. 84, no. 6, p. 066106, 2011.

[11] X. Ding and T. Jiang, "Spectral distributions of adjacency and laplacian matrices of random graphs," *The Annals of Applied Probability*, vol. 20, no. 6, 2010.

[12] P. Erdős and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.

[13] D. E. Fishkind, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe, "Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 23–39, 2013.

[14] Z. Füredi and J. Komlós, "The eigenvalues of random symmetric matrices," *Combinatorica*, vol. 1, no. 3, pp. 233–241, 1981. [Online]. Available: http://dx.doi.org/10.1007/BF02579329

[15] V. L. Girko, *Theory of Random Determinants*, ser. Mathematics and Its Applications. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1990.

[16] ——, *Theory of Stochastic Canonical Equations*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2001, vol. 1.

[17] S. Heimlicher, M. Lelarge, and L. Massoulié, "Community detection in the labelled stochastic block model," *arXiv preprint arXiv:1209.2910*, 2012.

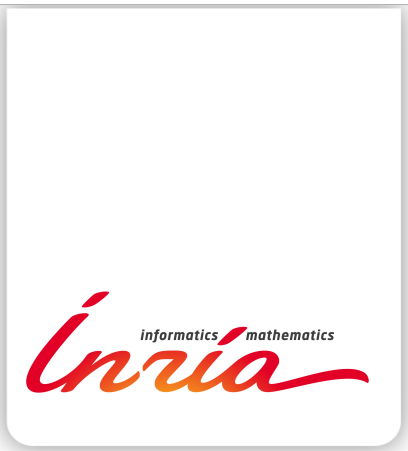
[18] R. R. Nadakuditi and M. E. Newman, "Graph spectra and the detectability of community structure in networks," *Physical review letters*, vol. 108, no. 18, p. 188701, 2012.

[19] Y. Saad, *Numerical methods for large eigenvalue problems*. SIAM, 1992, vol. 158.

[20] P. Van Mieghem, *Graph spectra for complex networks*. Cambridge University Press, 2011.

[21] V. Vu, "Spectral norm of random matrices," *Combinatorica*, vol. 27, no. 6, pp. 721–736, 2007. [Online]. Available: http://dx.doi.org/10.1007/s00493-007-2190-z

Contents



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399